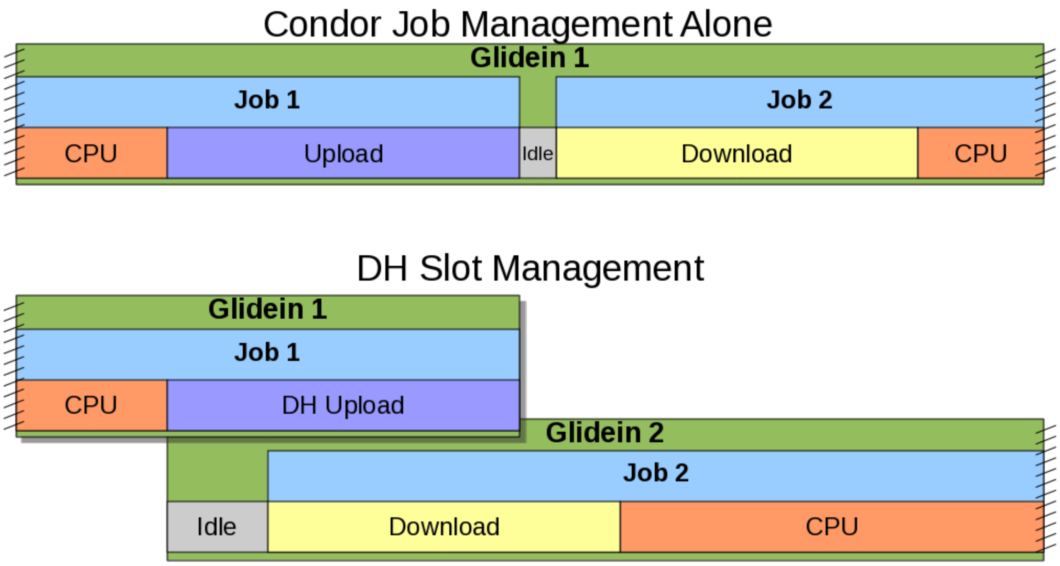
Slot Management for Data Return

Overview

The Data Handling (DH) Slot Management system replaces user-written, in-job upload systems. Instead of including the return mechanism into the executable, a job is submitted with extra information in the submit file. When the CPU intensive phase of the job is done, the job exits, and DH Slot Management begins. First, the system looks for files that match a regular expression filter. If the files are large enough, a new slot is created in the background, which can take new jobs. Meanwhile, a plugin is activated based on the data return system specified in the submit file, and the upload begins.



Drawing 1: The new system reduces wasted time by creating another glidein

Upload and download speeds do not compete locally due to duplexing. The network usage for simultaneous upload and download should be less than the maximum throughput of the hard disk system on the worker node. Overall bandwidth usage for the entire grid will not increase more than the throughput increase due to slot management. So, as long as the internet connection of the whole grid to the submitters is not at capacity, simultaneous network usage will improve throughput.

Ideally, CPU intensive phases should be adjacent, allowing the full resources of the system to be used continuously. DH slot management does not yet replace the download mechanism. It would be more difficult to support download methods and file selection. Since downloading is currently handled in the job, a the job would have to be suspended once downloading is complete or, the download system would have to be separated from the main job. Neither are viable options.

Evan Boldt 7/19/2010

Throughput Change

Definitions:

Run time:

The time for the job to finish. Includes the time to download data to the worker node and the CPU intensive phase, but does not include upload time

Upload Duration:

The time for data to be uploaded to its destination from the worker node.

New Slot Idle Time:

The time for a newly created glidein to be recognized by the collector and assigned a job. Expected between 10 – 60 seconds

Projected throughput change as a percentage:

(Upload Duration – New Slot idle Time) / (Run Time + Upload Duration)

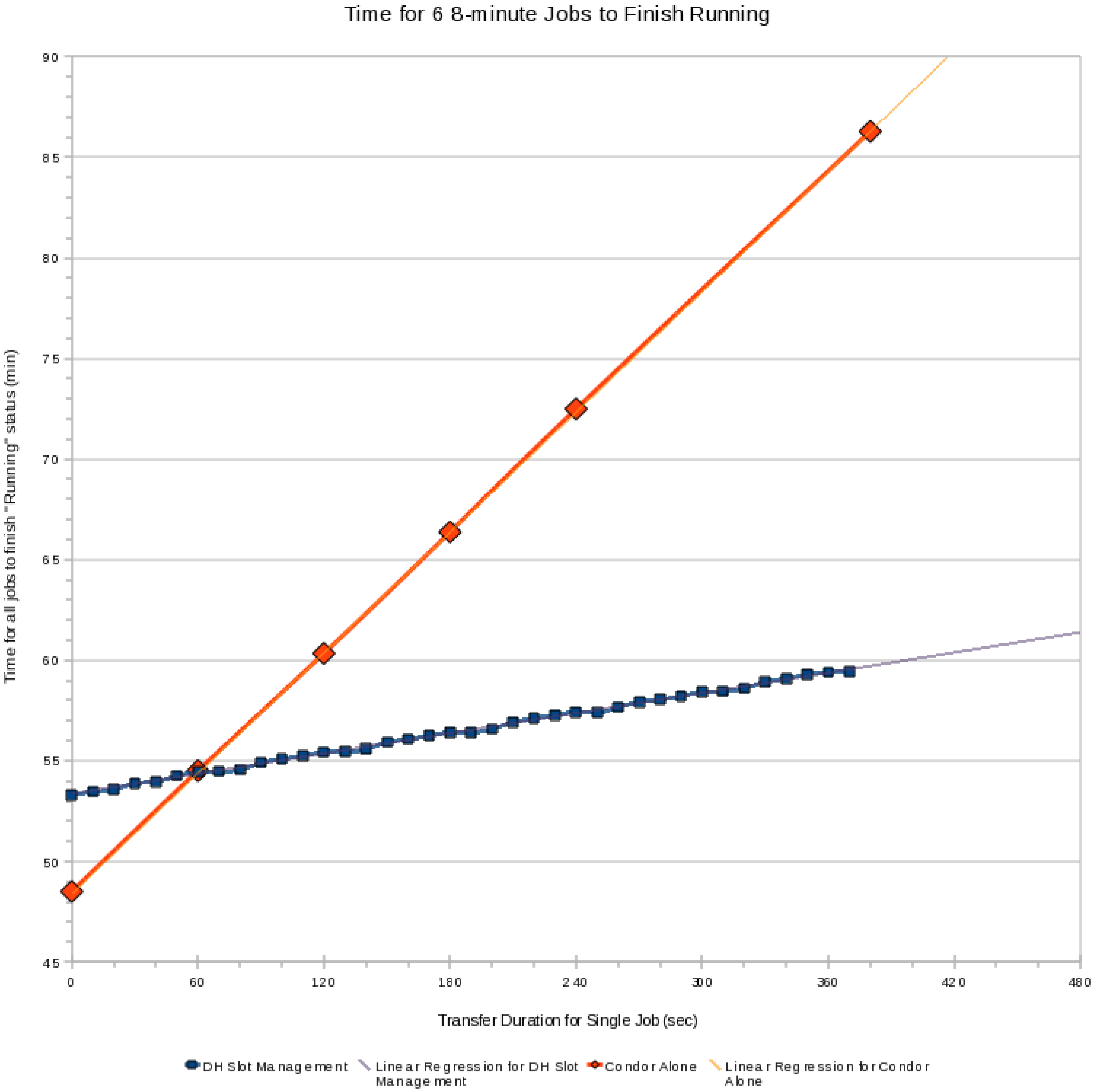
Upload Duration (minutes)	**Run Time (hours)**	**Idle Time (seconds)**	**Throughput Change (%)**
5	1	60	6.15%
10	1	60	12.86%
30	1	60	32.22%
5	3	60	2.16%
10	3	60	4.74%
30	3	60	13.81%
5	6	60	1.10%
10	6	60	2.43%
30	6	60	7.44%
		Max	32.22%
		Min	1.10%
		Average	9.21%

Table 1: Examples of projected throughput changes with estimated real-world values

While the maximum concurrent upload slots has not been reached, throughput should follow the formula. When the glidein hits the maximum number of data slots allowed, the benefit of overlapping transfer durations ceases. New slots are not created until the data is handled. The maximum should not be reached as long as upload duration does not exceed run time, which should not happen unless there is network failure. In the event of network failure, condor deletes the output data and gets a new job. Slot management allows jobs to remain in the worker node until the output is uploaded, further reducing waste.

Testing Data:

In this test, there are 6 jobs submitted in each pass, each job having an 8 minute run time. Each pass increases simulated upload durations. Testing was done on a single virtual machine, which ran the factory, frontend, and jobs locally. The factory submits only one glidein to the local machine.



A pass does not finish until all jobs have finished reporting "Running" status in condor_q. As a result, the run time for DH slot management increased by a factor of one transfer duration. This is because of the transfer for the last job. Even though another glidein was created, and could accept another job, there were no jobs remaining to accept. So, there is no throughput change if there are no jobs waiting in the queue.

While the run time for DH slot management increases by one upload duration for every pass, the run time for condor alone increases by six upload durations for every pass. However, since the DH slot management system is able to start a new job during the last transfer, the difference in run time per pass is six upload durations, assuming that more jobs are always available and neglecting new glidein idle time.

Since a glidein is ready at the start of each pass, there are 5 new slot idle times for each DH slot management pass. According to the pass with 0 upload duration, new slot idle time is approximately 60 seconds per slot or six minutes per pass. In cases where transfer time is less than new slot idle time, decreased performance is seen. If transfer delays are constant and are 60 seconds per job, and the transfer speed is two megabytes per second, then the total size of the transfers per job must be greater than 120 megabytes for a gain in throughput. DH Slot Management could attempt to guard against these cases, but it is difficult to reliably predict transfer duration. Theoretically, new slot idle time could be reduced by decreasing the communication interval between condor components, though it is unclear that this would be practical due to the extra network traffic.